analytical chemistry

pubs.acs.org/ac Article

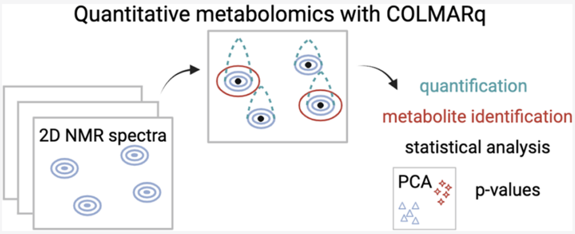
COLMARq: A Web Server for 2D NMR Peak Picking and Quantitative Comparative Analysis of Cohorts of Metabolomics Samples

Da-Wei Li, Abigail Leggett, Lei Bruschweiler-Li, and Rafael Brüschweiler*

Cite This: *Anal. Chem.* 2022, 94, 8674−8682 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Highly quantitative metabolomics studies of complex biological mixtures are facilitated by the resolution enhancement afforded by 2D NMR spectra such as 2D $^{13}C-^{1}H$ HSQC spectra. Here, we describe a new public web server, COLMARq, for the semi-automated analysis of sets of 2D HSQC spectra of cohorts of samples. The workflow of COLMARq includes automated peak picking using the deep neural network DEEP Picker, quantitative cross-peak volume extraction by numerical fitting using Voigt Fitter, the matching of corresponding cross-peaks across cohorts of spectra, peak volume normalization between different spectra, database query for metabolite identification, and basic univariate and multivariate statistical analyses of the results. COLMARq allows the analysis of cross-peaks that belong to both known and unknown metabolites. After a user has uploaded cohorts of 2D $^{13}C-^{1}H$ HSQC and optionally 2D $^{1}H-^{1}H$ TOCSY spectra in their preferred format, all subsequent steps on the web server can be performed fully automatically, allowing manual editing if needed and the sessions can be saved for later use. The accuracy, versatility, and interactive nature of COLMARq enables quantitative metabolomics analysis, including biomarker identification, of a broad range of complex biological mixtures as is illustrated for cohorts of samples from bacterial cultures of *Pseudomonas aeruginosa* in both its biofilm and planktonic states.


Quantitative metabolomics with COLMARq
2D NMR spectra
quantification
metabolite identification
statistical analysis
PCA
p-values

INTRODUCTION

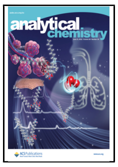
Metabolomics is the comprehensive identification and quantification of the small molecules involved in metabolic pathways in a biological system, known as metabolites.[1,2] Metabolites are the substrates and products of many biological processes; therefore, measuring the metabolic profile captures a snapshot of cellular activity. Metabolomics is also the most downstream omics strategy; therefore, it is influenced by upstream genetic and protein changes or environmental factors, making it uniquely reflective of the phenotype.[3] For these reasons, metabolomics approaches have proven valuable for diagnostics and monitoring of the treatment of a multitude of conditions and diseases, the characterization of regulatory biochemical processes, or applications in food science and nutrition.[4−6]

Intrinsic to the majority of successful metabolomics studies is the ability to accurately detect and quantify metabolites from a cohort of samples in a highly reproducible manner. Nuclear magnetic resonance (NMR) spectroscopy is a useful and powerful tool due to its inherent high reproducibility, resolution, and quantitative capabilities.[7−11] NMR is also nondestructive to the sample and does not require additional sample derivatization or separation steps, such as chromatography. NMR is uniquely adept for quantitative untargeted metabolomics because it can produce quantitative data for all reasonably abundant known and unknown metabolites present in a complex mixture in a single measurement.[8,12,13]

1D ^{1}H NMR is often utilized due to its short measurement time and quantitative nature. Several automated tools for 1D ^{1}H quantitative analysis have been developed such as MetaboLab,[14] BATMAN,[15] Bayesil,[16] AQuA,[17] ASICS,[18] and rDolphin.[19] However, complex mixtures containing metabolites with similar chemical motifs will cause peak overlap and crowded spectral regions, making metabolite identification ambiguous and quantification inaccurate.[10] These issues are largely resolved by collecting 2D NMR spectra, which adds additional resolution by correlating protons with neighboring nuclei such as carbon-13 or other protons.[13] Although 2D NMR spectra are not absolutely quantitative due to their dependence on J-couplings and differential spin relaxation times, peaks belonging to the same compound in spectra collected under the same parameters can be directly compared to determine relative concentrations for the quantitative determination of fold changes and statistical analysis between cohorts of samples.[10] If needed, absolute quantitation of 2D spectra can be achieved with the collection of reference spectra, spiking experiments, or specialty techniques like $HSQC_0$.[20]

The 2D $^{13}C-^{1}H$ HSQC offers significant resolution enhancement compared to 1D ^{1}H, ameliorating peak overlap. In

Received: February 23, 2022
Accepted: May 5, 2022
Published: June 7, 2022



addition, 2D $^1H-^1H$ TOCSY spectra aid in metabolite identification by providing intramolecular chemical connectivity information within spin systems of metabolites. This combined 2D HSQC and TOCSY approach, as implemented in our previously described COLMARm web server with its database of over 750 reference spectra, affords comprehensive, accurate, and efficient metabolite identification.[21] Still, extracting high-quality quantitative information from spectra remains a major challenge in NMR-based metabolomics.[22] The additional steps necessary for metabolite quantification, including peak picking, fitting, and matching, during the course of the analysis of cohorts of samples containing hundreds to thousands of peaks per spectrum can be ambiguous, time-consuming, and tedious.[10] A few tools have begun to take advantage of the increased resolution offered by 2D NMR to improve quantitative analysis of 1D ^1H spectra. Dolphin combines 1D ^1H and 2D J-resolved spectra to enhance reliability and accuracy of metabolite matching to reference spectra. In this method, 2D J-resolved spectra are used to identify targeted metabolites followed by quantification by lineshape fitting of the corresponding peaks in the 1D ^1H spectra. The user also has options for referencing and normalization, but quantification by this method is still limited by the extent of the 1D peak overlap.[22] In the R package *specmine*, 2D spectra are represented as a matrix, the dimensionality is reduced to a 1D *specmine* dataset to reduce computational cost, and then spectra can be plotted for visualization, peak detection, and measurement of peak intensities.[23] However, *specmine* requires coding experience (in R) and does not perform metabolite identification. Beyond these recent methods, there are no automated tools available for the identification and quantification of metabolites in 2D spectra and subsequent analysis.[13]

Here, we present the new public COLMARq web server, which facilitates the semi-automated, quantitative analysis of cohorts of 2D NMR spectra in an accurate and efficient manner (Supporting Information Figure S1). The COLMARq workflow (Figure 1) involves uploading of cohorts of 2D HSQC and 2D TOCSY spectra, peak picking, peak fitting, peak matching between samples, data normalization, database query, peak and metabolite-based statistical analysis, and data export of the results. This allows the user to easily input NMR spectra and efficiently arrive at quantitatively interpretable results such as *p*-values for metabolite concentration differences between groups or multivariate analysis. These tasks are performed in an automated manner while allowing for user input and manual correction as needed. After explaining the capabilities of COLMARq, it is demonstrated for a comparative quantitative analysis of cohorts of *P. aeruginosa* bacterial cultures in biofilm versus planktonic growth modes.

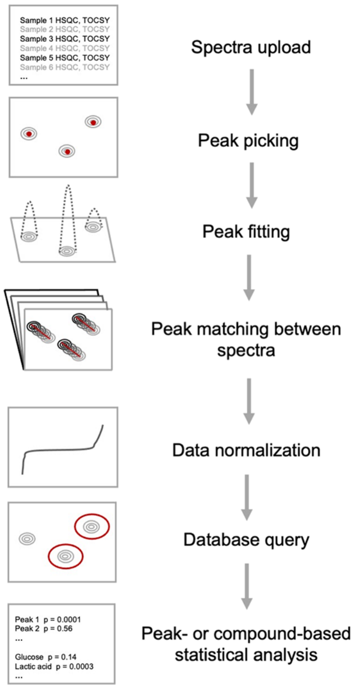


Figure 1. Workflow for the semi-automated quantitative analysis of HSQC spectra of metabolite mixtures by the COLMARq web server. COLMARq allows for upload of cohorts of HSQC and TOCSY spectra, automated peak picking, peak fitting for quantification, peak matching between spectra, data normalization via ratio analysis, database query for metabolite identification, and peak- and compound-based uni- and multi-variate statistical analyses.

EXPERIMENTAL SECTION

Sample Preparation. *P. aeruginosa* strain PAO1[24] cultures were grown overnight in lysogeny broth (LB) (Sigma Aldrich) and diluted to OD_{600} = 0.1. Then, cultures were scaled and grown planktonically in 50 mL of LB at 220 rpm at 37 °C for 24 h and as a biofilm on LB plates (28.4 cm^2) containing 1.5% (w/v) agar, statically, at 37 °C in 5% CO_2 for 48 h (n = 9) for metabolomics experiments.

Planktonic cultures were harvested by centrifugation at 4,300 × *g* for 20 min at 4 °C and washed with 1 mL of phosphate-buffered saline (PBS). Biofilm cultures were harvested by scraping with a sterile loop. Samples were immediately resuspended in 600 μL of cold 1:1 methanol (Fisher)/double distilled H_2O (ddH_2O) for quenching. Stainless-steel beads (SSB14B) (300 μL) (1.4 mm) were added, and cells were lysed using a Bullet Blender (24 Gold BB24-AU by Next Advance) at a speed of 8 for 9 min at 4 °C.[25] An additional 500 μL of 1:1 methanol/ddH_2O was added, and the sample was centrifuged at 14,000 × *g* for 10 min at 4 °C to remove solid debris. Methanol/ddH_2O/chloroform (Fisher) (1:1:1) was added for a total volume of 24 mL.[26,27] The sample was vortexed and centrifuged at 4,300 × *g* for 20 min at 4 °C for phase separation. The aqueous phase was collected, and the methanol content was reduced using rotary evaporation, followed by lyophilization overnight. For NMR measurements, the samples were resuspended in 200 μL of NMR buffer (50 mM sodium phosphate buffer in D_2O at pH 7.2 with 0.1 mM DSS (4,4-dimethyl-4-silapentane-1-sulfonic acid) for referencing) and centrifuged at 20,000 × *g* for 15 min at 4 °C for removal of any residual protein content. The pellet was washed with 100 μL of NMR buffer, and the supernatants were combined and transferred to a 3 mm NMR tube with a Teflon cap and sealed with parafilm.

NMR Experiments. NMR spectra were collected at 298 K on a Bruker AVANCE III HD 850 MHz solution-state spectrometer equipped with a cryogenically cooled TCI probe. 2D $^1H-^1H$ TOCSY spectra were collected (Bruker pulse program "dipsi2ggpphpr") with 256 complex t_1 and 2048 complex t_2 points for a measurement time of 4 h. The spectral widths along the indirect and direct dimensions were 10,202.0 and 10,204.1 Hz, and the number of scans per t_1 increment was 14. 2D $^{13}C-^1H$ HSQC spectra (Bruker pulse program "hsqcetgpsisp2.2") were collected with 512 complex t_1 and 2048 complex t_2 points for a measurement time of 16 h. The spectral widths along the indirect and direct dimensions were

https://doi.org/10.1021/acs.analchem.2c00891

34,206.2 and 9375.0 Hz, and the number of scans per t_1 increment was 32. The transmitter frequency offset values were 75 ppm in the ^{13}C dimension and 4.7 ppm in the ^1H dimension for all experiments. NMR data was zero-filled four-fold in both dimensions, apodized using a cosine-squared window function, Fourier transformed, and phase corrected using NMRPipe.[28]

■ RESULTS

The individual steps are listed in the flowchart of COLMARq (Figure 1), and they are explained in more detail in the following.

Since most metabolomics studies typically start out with metabolite identification, COLMARq was designed to work directly with the results of previous COLMARm session(s) used for metabolite identification. Hence, for each sample, the processed 2D HSQC and optionally TOCSY NMR spectra in the frequency domain are first uploaded to the COLMARm web server, followed by peak deconvolution and spectral referencing (if necessary). It accepts the spectral data formats of Bruker Topspin (ASCII), Mnova, NMRPipe, and Sparky. If the user has prior knowledge of the metabolite composition of the samples and is familiar with the functions of COLMARm, the spectral files can also be directly uploaded to COLMARq in batch mode. First, all cross-peaks are identified by automated peak picking, which is critical for all subsequent steps. COLMARq and COLMARm support two types of peak pickers: the default method is our recently introduced deep neural network DEEP Picker, which has proven highly effective for crowded 2D spectra of proteins and metabolomics samples.[29] As an alternative, a traditional peak picker can be selected, which is based on a Laplacian spectral filter amplifying shoulder peaks at the cost of increased noise and some false positive peak identification in highly crowded regions. Our traditional peak picker is similar to existing peak pickers implemented in Mnova and other tools.[30,31]

Next, each identified cross-peak is quantified for the purpose of determining the relative concentration of the metabolite it belongs to. This is accomplished by numerical fitting of the cross-peaks using the software "Voigt Fitter" specifically developed for this task. After appropriate apodization using a cosine square or 2π-Kaiser window function, NMR lineshapes follow in good approximation Voigt profiles, which are hybrids between Lorentzian and Gaussian profiles, along both frequency dimensions. Each 2D HSQC cross-peak is characterized by seven parameters: the peak position along each dimension (which can be off the underlying digital spectral grid), the peak amplitude or volume, and the peak shape, whereby the peak shape is determined by its two Voigt parameters along each dimension. For many $^{13}C-^1H$ HSQC spectra in metabolomics, the cross-peaks have in good approximation a Gaussian shape and thus can be fitted with only five parameters. Using the output of the peak picker as initial values for fitting, the Voigt Fitter performs a non-linear least-squares fit to simultaneously optimize peak parameters of all peaks to reproduce the original spectrum. While nonoverlapping peaks can be fitted individually quite efficiently, fitting of large overlapping peak clusters requires simultaneous fitting of N cross-peaks identified in the cluster. Most non-linear least square fitting algorithms, such as the Levenberg−Marquardt algorithm and its derivatives,[32] involve the iterative diagonalization of a 5N × 5N square matrix, which computationally scales with $O(N^3)$. As a consequence, for sizable N, the fitting process can become very slow, even on modern computer workstations. To address this issue, we implemented in the Voigt Fitter software a Gaussian mixture-type model algorithm,[33] which scales linearly with N, allowing the rapid fitting of complex spectra with an essentially unlimited number of both overlapping and non-overlapping cross-peaks as typically encountered in metabolomics spectra. The Gaussian mixture-type model algorithm solves the problem iteratively where each iteration includes the following steps: (1) calculate the theoretical spectrum of each peak using its current peak parameters; (2) aggregate the theoretical spectra of all peaks to obtain the total theoretical spectrum; (3) calculate for each individual peak the ratio of its (theoretical) spectrum and the total (theoretical) spectrum; (4) deconvolute the experimental spectrum into the spectra of individual peaks in a way such that the ratio of individual spectral peaks and the total experimental spectrum is the same as the ratio obtained in step 3; and (5) fit each peak using the deconvoluted spectrum as a starting point and update peak parameters. The algorithm will go back to step (1) until the change of the peak parameters falls below a predefined cutoff. In step (5) of the algorithm, the nonlinear least squares fit is performed sequentially for each individual cross-peak in a five-dimensional parameter space (in the case of Gaussian peak shapes), rather than simultaneously for all N cross-peaks in a 5N-dimensional parameter. Hence, the computational effort of the algorithm scales linearly with N, i.e., O(N), allowing a dramatic speed-up in the fitting of spectra with large numbers of peaks as typically encountered in metabolomics applications. In contrast to other fitting software, Voigt Fitter does not require the selection of spectral subregions for efficient fitting as it can autonomously handle entire spectra with several thousand cross-peaks. Voigt Fitter also does not require the manual addition or elimination of peaks for improved fitting as DEEP Picker reliably produces a high-quality set of cross-peaks, including their positions, lineshapes, and amplitudes, as a starting point for Voigt Fitter. As a benchmark, the fitting of the 1772 cross-peaks of a 2D $^{13}C-^1H$ HSQC spectrum of the *P. aeruginosa* biofilm, where the largest peak cluster contains 142 peaks, takes only about 20 s. By contrast, due to its unfavorable scaling property, a traditional non-least-squares fitting approach takes many hours or even days. An illustration of a complex spectral region of the biofilm spectrum and its fitted counterpart is shown in Figure 2, demonstrating the high accuracy of the Voigt Fitter even for highly overlapped cross-peak clusters.

The next step in the COLMARq workflow is to match peaks stemming from the resonance signal of a certain spin of the same metabolite across the entire batch of spectra. The peak matching algorithm takes into account peak positions (chemical shifts), peak heights, possible peak multiplets due to scalar J-couplings, and peak picking consistency among different spectra. In metabolomics samples, the vast majority of cross-peaks have well-defined positions that remain essentially unchanged from sample to sample. However, a small number of cross-peaks can move by as much as 0.02 or 0.2 ppm along the proton or carbon dimension, respectively. This can be caused by slight variations of sample conditions among replicates, such as alterations in pH. Besides chemical shift information, the peak matching algorithm also takes into account peak amplitudes. Specifically, peaks whose amplitudes are within a factor 10 of each other in different samples are preferred for matching. If this is not possible within the chemical shift cutoff, the peak matching algorithm will then try to match peaks with amplitude ratios exceeding 10.

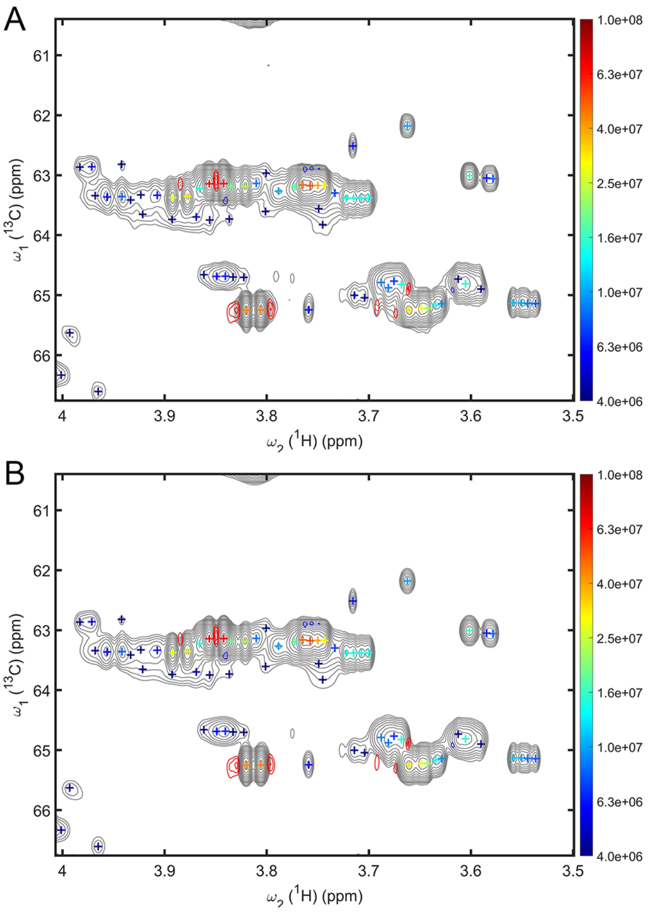


Figure 2. Selected region of the ^{13}C−^{1}H HSQC spectrum of the biofilm (A) and the reconstructed spectrum by COLMARq (B) from the fitted peaks. Contour lines are plotted using a logarithmic scale and the fitted cross-peaks are indicated by plus symbols that are color coded according to the cross-peak amplitudes (logarithmic scale, see color sidebar). Residual fitting errors are plotted in both panels (A) and (B) as red (positive) and blue (negative) contour lines using the same scale.

Spectral multiplets observed in metabolomics HSQC spectra should be matched as a group against the same kind of multiplets in other samples. An example of matched doublets is shown in Figure 3 (blue cluster). For low sensitivity multiplets (with amplitudes smaller than 10 times the noise level) or multiplets that strongly overlap with other peaks, DEEP picker (and Voigt fitter) may interpret the same feature as a multiplet in some samples and as a single peak in others. An example of such a case is also shown in Figure 3 (red cluster) where the consensus peak was identified as three peaks in Samples #0, #2, and #3 and four peaks in Sample #1. While the peak matching algorithm will assign a lower confidence score to these types of imperfect matching results, they can still be useful for downstream analysis. Because of the sometimes difficult and ambiguous nature of peak matching, it is recommended that the user check the peak matching results using the visualization plots on the web server to ensure the most accurate downstream quantitative analysis. The web server was designed with a high level of flexibility, allowing users to interactively make manual adjustments to the peak matching result. Based on the user's assessment of the confidence in the matching results of individual peaks, they can be adjusted or discarded during a later stage of the analysis.

Normalization of spectra is important to correct for variations in the total sample amount or overall sample concentration between replicates or cohorts, which may occur during sample collection, sample preparation, or data acquisition.[34−36] For solution NMR-based studies in which the total volume of each sample can be controlled during sample preparation, the potential global dilution factor for each sample should be accounted for during data analysis. For this purpose, COLMARq supports the widely used median fold change method,[34] which works well when many metabolites have a similar concentration across all samples. This method determines the median fold change between samples as a robust estimate of the dilution factors between samples. Specifically, the COLMARq normalization tool estimates the normalization factors between a reference sample specified by the user and all other samples. For each pair of samples, the tool calculates the fold-change ratio of all matched peaks, rank orders the ratios, and then uses the mean of the median 30% fold-change ratios as the normalization factor. The accuracy of this approach depends on the quality of peak matching in the previous step. As mentioned above, the COLMARq server gives users the option to manually adjust peak matching and exclude matched peaks that have a low confidence score.

As an example, we uploaded cohorts of nine spectra from *P. aeruginosa* planktonic and biofilm cultures to COLMARq for statistical analysis. A screenshot of the normalization plot of the web server is displayed in Figure 4A. Peak volumes of Sample #3 were divided by the corresponding peak volumes of Sample #2, which was chosen as the reference spectrum, and the resulting peak volume ratios were rank ordered. Figure 4A shows the logarithm of the ratios vs peak number, giving rise to a characteristic rotated sigmoidal curve. The tails on both ends show peaks that mostly differ between samples (smallest ratios on the left and largest ratios on the right), while the relatively flat center reflects that the dilution factor between the samples is minimal. Averaging the median 30% of ratios results in a normalization factor of 1.005 for this sample, indicating that any dilution effect for Sample #3 vs Sample #2 is minimal. This type of normalization plot can be generated for each sample to obtain a visual impression of potential dilution effects and determine whether the underlying assumption of this method is valid, namely, that the majority of ratios between metabolites are, in good approximation, constant as manifested in a flat middle range of the rotated sigmoidal curve. The peak volumes of each spectrum are then divided by the normalization factor to make them quantitatively comparable to the reference spectrum and to each other for subsequent statistical analysis.

Once cross-peaks are quantified and matched across all samples, statistical analysis can be performed in a standard manner. Although statistical analysis tools are widely available, the COLMARq server also provides limited univariate and multivariate statistical analysis capabilities to readily give users information about cross-peaks or metabolites that have statistically significant concentration differences between cohorts. The user can sort the uploaded samples into two groups with the option to selectively exclude samples from statistical analysis. At this time, COLMARq provides peak-based *p*-value analysis (*t* test), including a histogram of all *p*-values. For the *p*-value calculation, the two cohorts are assumed to be both normally distributed with equal variance. COLMARq also allows users to perform a peak-based principal component analysis (PCA) as an unsupervised multivariate statistical analysis method commonly used in metabolomics for the visual clustering of samples in a score plot based on the covariation of cross-peak volumes to assess separation between cohorts.

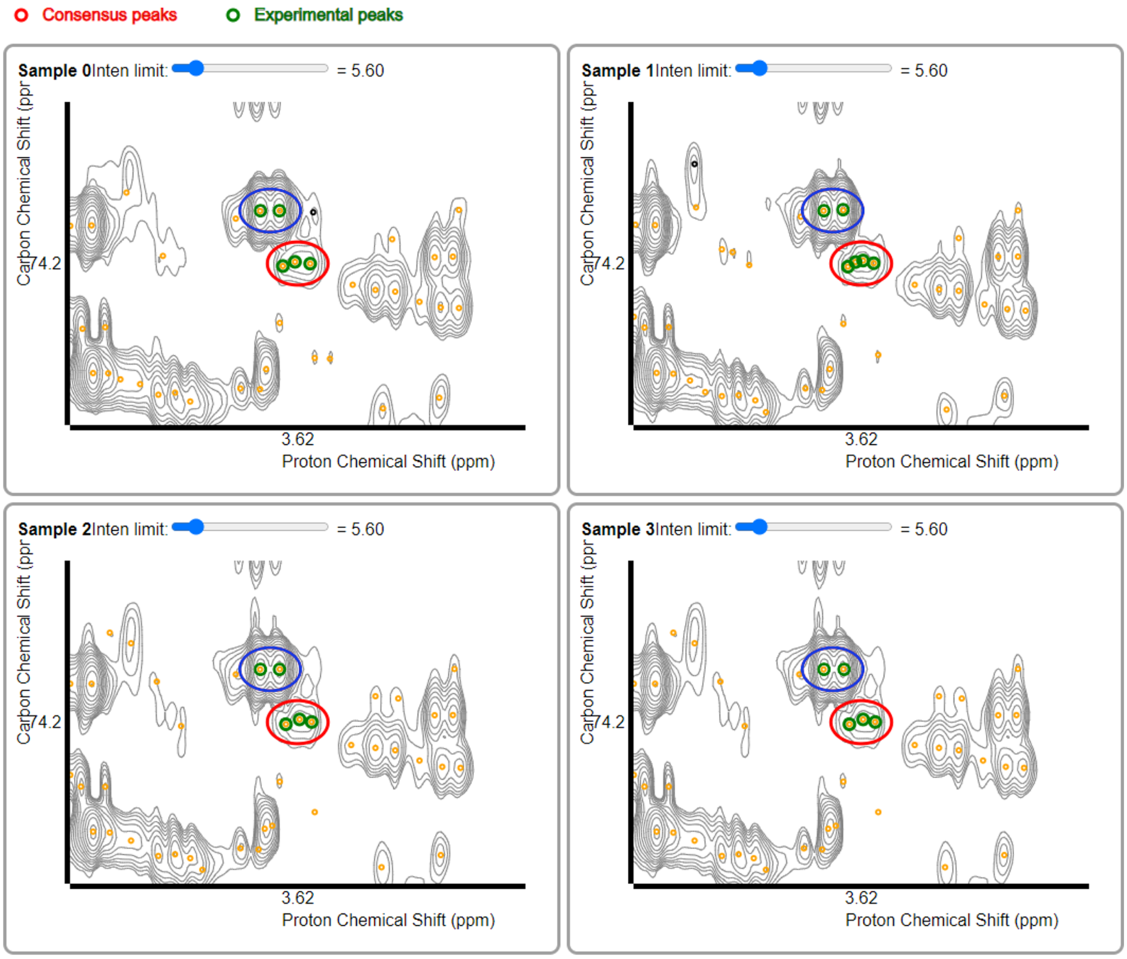

○ Consensus peaks
○ Experimental peaks
Sample 0 Inten limit: = 5.60
Sample 1 Inten limit: = 5.60
Sample 2 Inten limit: = 5.60
Sample 3 Inten limit: = 5.60
Carbon Chemical Shift (ppm)
74.2
3.62
Proton Chemical Shift (ppm)

Figure 3. Example of two matched (consensus) cross-peaks across the $^{13}C-^{1}H$ HSQC spectra of four different *P. aeruginosa* samples. A high sensitivity doublet is labeled by blue ellipses whereas a low sensitivity multiplet is labeled by red ellipses containing either 3 or 4 individual cross-peaks across the different spectra. Individual peaks that belong to these two consensus peaks are labeled as green circles. All other peaks that were part of other consensus peaks are labeled as small orange circles, and nonconsensus peaks that appear only in one spectrum are labeled as black circles.

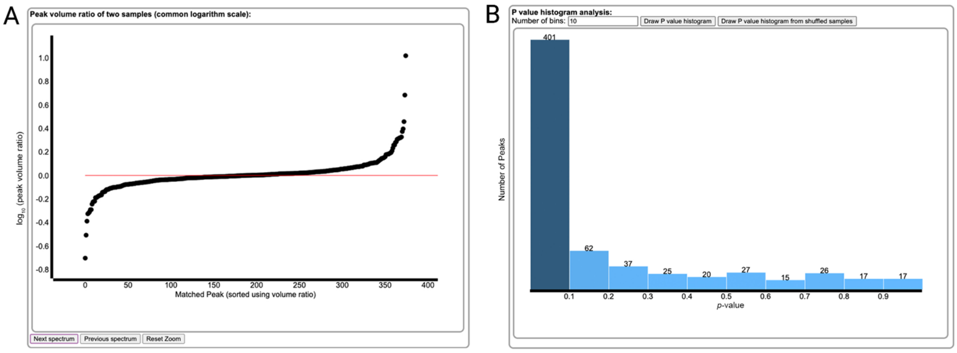
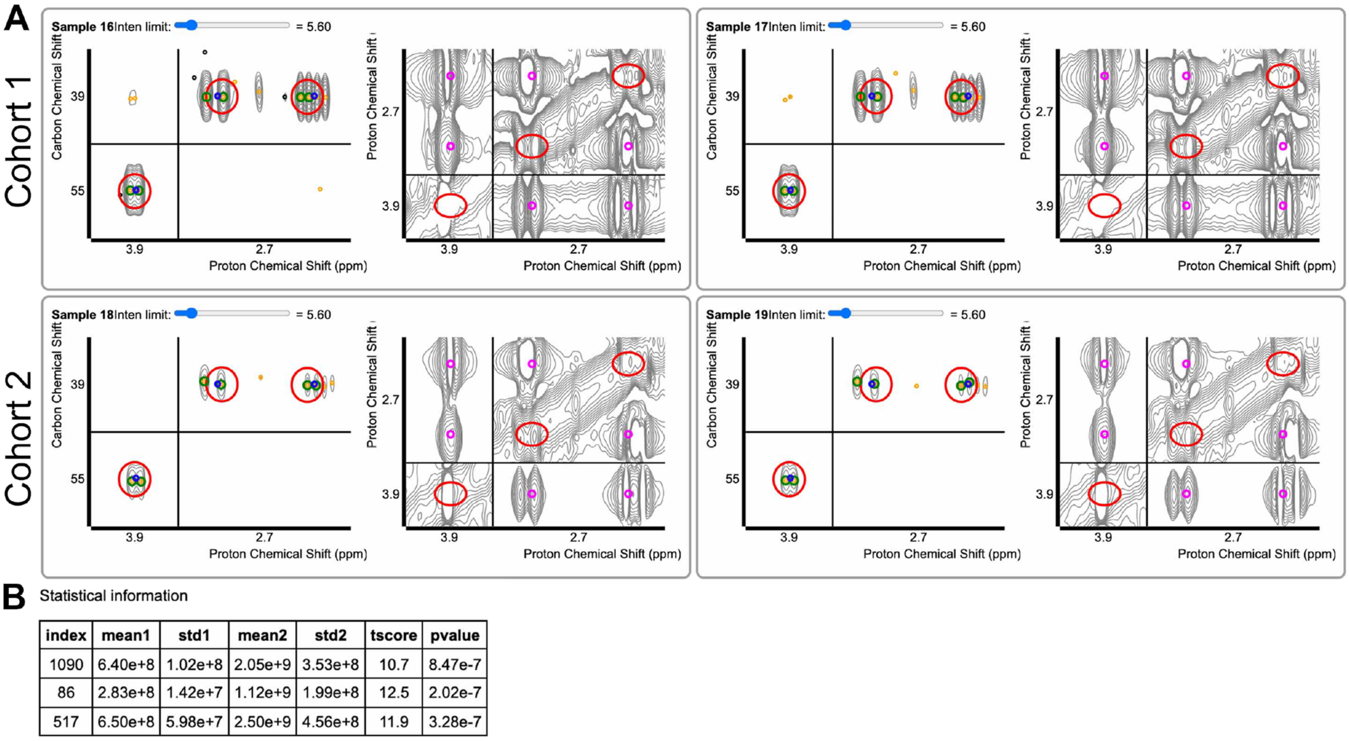


Figure 4. COLMARq can (optionally) perform data normalization and peak-based statistical analysis. For normalization, peak volumes are divided by matched peaks of a user-selected reference spectrum and the $\log_{10}$(ratios) are rank ordered and plotted versus the number of peaks (A). The average ratio of the flat central part, calculated as the median of the 35–65% percentile ratios, determines the normalization factor for each spectrum and all peaks are divided by this factor. After peak-based statistical analysis, COLMARq displays the *p*-value histogram (B) showing the distribution of *p*-values from *t*-tests. In this example, a high number of significant differences between cohorts reflect the inherent metabolic heterogeneity of the *P. aeruginosa* planktonic and biofilm cultures.

For metabolites and peaks that are observable in some samples and unobservable in others, it is generally useful to set the missing amplitudes to either 1/2 or 1/3 (default) of the detection limit of the experiment, rather than setting them to zero. In COLMARq, the peak amplitude detection limit can be defined by the user as a fixed multiple of the noise level automatically determined for each spectrum. For this purpose, from all observable peaks, an empirical relationship between peak volumes and peak amplitudes is established, which is then used to estimate the peak volume of peaks with amplitudes at 1/3 of the peak height detection limit.


A
Cohort 1
Cohort 2
Sample 16 Inten limit: = 5.60
Sample 17 Inten limit: = 5.60
Sample 18 Inten limit: = 5.60
Sample 19 Inten limit: = 5.60
Carbon Chemical Shift
Proton Chemical Shift
39
55
2.7
3.9
Proton Chemical Shift (ppm)
B

Statistical information

index	mean1	std1	mean2	std2	tscore	pvalue
1090	6.40e+8	1.02e+8	2.05e+9	3.53e+8	10.7	8.47e-7
86	2.83e+8	1.42e+7	1.12e+9	1.99e+8	12.5	2.02e-7
517	6.50e+8	5.98e+7	2.50e+9	4.56e+8	11.9	3.28e-7

Figure 5. The COLMARq user interface enables visual inspection of the metabolite matches after database query for metabolite identification. The user can click through each metabolite match for visual inspection of the specific spectral regions of both the HSQC (left) and TOCSY (right) spectrum of each sample for judging the match (A). Panel (A) shows the spectra of two representative samples of each cohort matched to aspartate for Cohort 1 in the top row and Cohort 2 in the bottom row. In the HSQC spectra, the blue circles represent the expected database peaks for a metabolite, the red circles indicate the consensus peak position from user spectra for each expected metabolite peak, and the green circles mark the peaks the user selects for quantification, which can be manually edited. In the TOCSY spectra, the pink circles mark expected cross-peaks for a metabolite match. With its high peak matching ratio in the HSQC and the presence of the expected TOCSY cross-peaks, aspartate is a good match. The chart shown in panel (B) reports the quantitative information for each peak of the metabolite match, denoted by its unique peak index, including mean values of peak volumes with their standard deviations for each cohort along with *t*-scores and *p*-values from *t* tests.

In our demonstration with *P. aeruginosa*, a total of 1302 distinct cross-peaks were picked in each spectrum with 782 peaks showing a significant difference between cohorts with $p <$ 0.05. A screenshot of the *p*-value histogram from the web server (Figure 4B) including only peaks present in all 18 spectra shows a substantial number of cross-peaks whose volumes systematically differ between cohorts (histogram bar on very left) reflecting the inherent metabolic heterogeneity of the *P. aeruginosa* planktonic and biofilm cultures. Of the significantly different peaks, 493 do not match to any known metabolites in the database, highlighting the potential of peak-based statistical analysis for the characterization also of unknown metabolites.

COLMARq also provides metabolite database query capabilities directly adopted from COLMARm. If an experimental consensus peak is within the predefined frequency cutoff of a database peak, it is classified as a "matched peak". The "matching ratio" is then defined as the ratio of the number of matched peaks to the total number of peaks of the database compound.[21] The default cutoff parameters for the ^{1}H and ^{13}C chemical shift differences are set at 0.04 and 0.4 ppm, respectively, and the lowest accepted matching ratio is set to 0.6. Users can alter these three parameters on the web server interface and repeat database query to see how they affect the returned matched metabolite list. If needed, users also have the option to interactively edit the cross-peaks matched to each metabolite database peak by drag and drop.

COLMARq aims at detecting all possible metabolite matches, whereby user visualization plays an important part to narrow down and confirm the true matches. In our demonstration with *P. aeruginosa* using cutoff parameters of 0.3 ppm for ^{13}C and 0.03 ppm for ^{1}H with a peak matching ratio of 0.6, a total of 169 metabolites were matched to the spectra. After manual editing, 66 metabolites were determined to be highly confident hits, marked as good or fair, and quantified. The total matched compound list included 68 tentative hits that were matched due to a peak overlap between similar metabolites but do not contain unique peaks. This can occur in highly crowded spectral regions pertaining to highly similar compounds such as carbohydrates and nucleotides, which comprise 47 of the 68 tentative hits. An additional 22 compounds were matched, but because they were present at low abundance with weak and missing peaks in many spectra, they were not quantified. If desired, the user can set stricter cutoff parameters to reduce the number of incomplete matches. Figure 5A shows a screenshot from the web server as an example of four interactive HSQC and TOCSY plots zoomed in on a metabolite match. The blue circles mark the expected cross-peak positions for this metabolite from the database, and the pink circles in the TOCSY mark expected TOCSY cross-peaks. As previously mentioned, the user can drag and drop the blue circles to select which experimental peaks are the best match for this metabolite. Another example of metabolite matching with more samples is shown in Supporting Information Figure S3.

In addition to the cross-peak based *p*-value analysis of Figure 4B, users can also perform compound-based *p*-value calculations. In this case, the relative concentration of a compound is calculated from the weighted average peak volume over all its cross-peaks. By default, all cross-peaks have the same weight, but users have the option to adjust the weights. For example, users can assign lower or even zero weight to weak peaks so that the relative concentration is dominated by the strongest and, hence, most quantitative peaks of a metabolite. Using a weight of zero to exclude peaks is useful in the case that one or more peaks belonging to a metabolite are overlapped with a peak from another metabolite allowing the inclusion of only unique peaks for accurate quantification. For *P. aeruginosa,* of the total 66 matched metabolites, 52 display a significant concentration difference between cohorts ($p < 0.05$). Figure 5B shows a chart with statistical information for an example metabolite match. The chart includes for each peak the mean and standard deviation for each cohort and the *t*-score and *p*-value between cohorts.

The COLMARq server provides several flexible options for the user to download both intermediate and final results for subsequent use. For example, users have the option to download the matched peak list with peak volumes in text format so that they can be used as input for further statistical analysis using the user's preferred software. Users can also download numerical peak-based or compound-based *p*-value results.

■ DISCUSSION

The high complementary of NMR to mass spectrometry makes NMR a powerful method for the targeted and untargeted quantitative analysis of metabolomics samples.[8] Due to NMR's unique versatility, it is not a surprise that there exist a variety of different NMR approaches, each with its own pros and cons. High-throughput applications involving large cohorts of samples typically rely on 1D ^{1}H experiments as it requires measurement times of only around 15 min per sample. On the flip side, the ability to uniquely identify a large number of metabolites from 1D spectra alone is limited due to crowded spectral regions that are difficult to deconvolute. In addition, strong peak overlap and background signals can compromise quantitation of individual peaks. It is therefore common to assist 1D NMR-based metabolomics studies with a very small number of 2D NMR experiments of selected samples for the verification of metabolite assignments.[37] 2D NMR spectra, such as ^{13}C−^{1}H HSQC, ^{1}H−^{1}H TOCSY, or ^{1}H−^{1}H COSY, provide vast resolution enhancement over 1D. However, the collection of 2D NMR spectra for samples that are limited by sensitivity, rather than the sampling of the indirect time domain, is typically associated with significantly prolonged measurement times. This applies in particular to ^{13}C−^{1}H HSQC spectra at ^{13}C natural abundance. At the same time, the first-rate resolution properties make them particularly well suited for semi-automated, quantitative analysis. For other 2D NMR spectra, nonuniform sampling along the indirect dimension or ultrafast 2D NMR can provide a significant speed-up over the traditional 2D NMR acquisition method.[38] Compared to 1D NMR, 2D NMR-based metabolomics is more involving during the NMR data acquisition stage. On the other hand, the 2D method offers a substantial time gain together with higher accuracy during the analysis part of a project.

Despite their widespread use for resonance assignment and metabolite identification purposes, it is still very uncommon to use 2D NMR spectra for fully quantitative metabolomics analysis. A number of standalone software has been introduced for the quantitative analysis of 2D NMR cross-peaks by peak fitting, including NMRPipe,[28] FMLR,[39] PINT,[40] INFOS,[41] and FitNMR,[42] using a range of different models for the peak shapes from Gaussian to lineshapes directly mirroring the apodization function used. These software programs have not been designed for the typical metabolomics workflow involving cohorts of complex spectra from different samples that require peak matching, which may explain their lack of routine usage in metabolomics. COLMARq offers a convenient integration by directly using metabolite assignments (from COLMARm) for quantification of cohorts of spectra, peak matching, normalization, and statistical analysis. COLMARq is the first publicly available web server to facilitate metabolite identification and fully quantitative analysis of 2D NMR spectra for metabolomics.

^{13}C−^{1}H HSQC spectra have a very clean baseline void of a background signal in most regions, which makes them particularly suitable for the highly quantitative analysis of a large number of peaks. COLMARq is best used in combination with COLMARm, where for each sample a COLMARm analysis for metabolite query is performed first. This is followed by the simultaneous uploading of all COLMARm sessions into COLMARq for quantification. For experienced users, the COLMARm upload step can be circumvented, and the spectra corresponding to all samples can be uploaded in batch mode to COLMARq for analysis. Normalization of peak volumes from different spectra is known to be important. The median ratio method implemented in COLMARq assumes that the concentration of a majority of metabolites remains unchanged, giving rise to the flipped sigmoidal profile of the rank-ordered ratios with an extended flat part in the middle percentile range. The graphical representation of this relationship by COLMARq (Figure 4A) allows the user a quick assessment whether this assumption is fulfilled and the normalization procedure is appropriate for a particular study.

The analysis of a large number of cross-peaks across a cohort of samples afforded by 2D NMR-based metabolomics also allows a meaningful analysis of the *p*-value distribution in the form of a histogram (Figure 4B). For two sample cohorts that are statistically indistinguishable, the *p*-value histogram should be flat, i.e., each *p*-value from 0 to 1 has the same probability.[43] Therefore, the *p*-value histogram provides a straightforward visual assessment whether the two cohorts are inherently different in their metabolomic makeup. This is particularly useful for pilot studies based on a relatively small number of samples to decide whether a larger scale study, for example, for the characterization of putative biomarkers, is warranted. A key advantage of 2D NMR-based metabolomics is that it works equally well for targeted and untargeted studies, including biological samples that are not commonly studied, involving potentially large numbers of cross-peaks belonging to both known and unknown metabolites. Such high-quality information is harder to obtain from 1D NMR-based metabolomics unless the metabolite composition, as for example for human serum, is mostly known.

COLMARq is largely automated by taking advantage of the very accurate peak identification performance by DEEP Picker as input for Voigt Fitter for quantification. In addition, manual editing is made possible, which is the most useful for peak matching between multiple spectra in strongly overlapped regions that show variations in peak positions between samples or for weak peaks that only show up in subsets of spectra. As a demonstration, we used COLMARq for the efficient, semi-automated analysis of metabolite extracts from cohorts of nine *P. aeruginosa* planktonic and biofilm cultures each. With over 32,000 spectral cross-peaks to analyze across all 18 2D HSQC spectra, manual analysis is tedious and can take months. Batch uploading of the 18 sets of 2D HSQC and TOCSY spectra to COLMARq (~30 min), automatic peak picking, fitting, and matching between spectra (~2.5 h), metabolite query against the database (few seconds), and normalization and statistical analysis (few seconds) were completed with the COLMARq server in only about 3 h. Manual adjustment of the automated peak matching between spectra to ensure accurate selection of peaks within multiplets and between samples is the most time-consuming step when working with a larger volume of samples. COLMARq provides visualization of all matched peaks and metabolites for a user-friendly approach to inspection and judgment of matches. The highly interactive nature of the web server facilitates simple adjustments during the course of all analysis steps. The user can go from the collected NMR spectra to a list of metabolites with their fold-changes, p-values, and a PCA plot between hours to a few days, depending on the number of samples and amount of manual adjustments required. In the *P. aeruginosa* samples, 66 metabolites were judged as good or fair database matches and 52 of these metabolites showed a significant difference between cohorts ($p < 0.05$). For a recent study, these results were exported and the metabolites were mapped to metabolic pathways to provide information about the differential metabolism of *P. aeruginosa* in the two growth modes.[44] COLMARq is not limited by sample type and therefore should be useful for the analysis of a wide variety of metabolomics applications.

In summary, the main goal behind the new COLMARq web server is to provide users a simple, intuitive, and versatile peak picking, fitting, and matching tool for a widest possible range of NMR-based metabolomics studies that is publicly accessible. The quantification, matching, and assignment of all peaks from the sample cohorts represent a comprehensive and fully quantitative approach for the downstream analysis in both targeted and untargeted metabolomics studies. COLMARq allows users to take full advantage of the resolution and quantitative power of 2D NMR-based metabolomics measurements, considerably facilitating the accurate, semi-automated, and efficient analysis of metabolomics data.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.analchem.2c00891.

Screenshots of (i) the COLMARq web server homepage and (ii) the user interface for visual inspection of either metabolite or cross-peak matching for a cohort of samples (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Rafael Brüschweiler − *Campus Chemical Instrument Center, Department of Chemistry and Biochemistry, and Department of Biological Chemistry and Pharmacology, The Ohio State University, Columbus, Ohio 43210, United States;* orcid.org/0000-0003-3649-4543; Email: bruschweiler.1@osu.edu

Authors

Da-Wei Li − *Campus Chemical Instrument Center, The Ohio State University, Columbus, Ohio 43210, United States;* orcid.org/0000-0002-3266-5272

Abigail Leggett − *Department of Chemistry and Biochemistry and Ohio State Biochemistry Program, The Ohio State University, Columbus, Ohio 43210, United States*

Lei Bruschweiler-Li − *Campus Chemical Instrument Center, The Ohio State University, Columbus, Ohio 43210, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.analchem.2c00891

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We thank Dr. Alexandar L. Hansen for valuable discussion about Voigt Fitter. This work was supported by the National Institutes of Health (R35GM139482 to R.B.). All NMR experiments were performed at the CCIC NMR facility at The Ohio State University.

■ REFERENCES

(1) Lindon, J. C.; Holmes, E.; Bollard, M. E.; Stanley, E. G.; Nicholson, J. K. *Biomarkers* **2004**, *9*, 1−31.

(2) Klassen, A.; Faccio, A. T.; Canuto, G. A. B.; da Cruz, P. L. R.; Ribeiro, H. C.; Tavares, M. F. M.; Sussulini, A., Metabolomics: Definitions and Significance in Systems Biology. In *Metabolomics: From Fundamentals to Clinical Applications*, Sussulini, A., Ed. Springer International Publishing: Cham, 2017; pp. 3−17, DOI: 10.1007/978-3-319-47656-8_1.

(3) Fiehn, O. *Plant Mol. Biol.* **2002**, *48*, 155−171.

(4) Holmes, E.; Wilson, I. D.; Nicholson, J. K. *Cell* **2008**, *134*, 714−717.

(5) Wishart, D. S. *Nat. Rev. Drug Discovery* **2016**, *15*, 473−484.

(6) LeVatte, M.; Keshteli, A. H.; Zarei, P.; Wishart, D. S. *Lifestyle Genomics* **2022**, *15*, 1−9.

(7) Nagana Gowda, G. A.; Raftery, D. *J. Magn. Reson.* **2015**, *260*, 144−160.

(8) Markley, J. L.; Brüschweiler, R.; Edison, A. S.; Eghbalnia, H. R.; Powers, R.; Raftery, D.; Wishart, D. S. *Curr. Opin. Biotechnol.* **2017**, *43*, 34−40.

(9) Bingol, K.; Bruschweiler-Li, L.; Li, D.; Zhang, B.; Xie, M.; Brüschweiler, R. *Bioanalysis* **2016**, *8*, 557−573.

(10) Crook, A. A.; Powers, R. *Molecules* **2020**, *25*, 5128.

(11) Edison, A. S.; Colonna, M.; Gouveia, G. J.; Holderman, N. R.; Judge, M. T.; Shen, X.; Zhang, S. *Anal. Chem.* **2021**, *93*, 478−499.

(12) Bingol, K.; Brüschweiler, R. *Anal. Chem.* **2014**, *86*, 47−57.

(13) Emwas, A. H.; Roy, R.; McKay, R. T.; Tenori, L.; Saccenti, E.; Gowda, G. A. N.; Raftery, D.; Alahmari, F.; Jaremko, L.; Jaremko, M.; Wishart, D. S. *Metabolites* **2019**, *9*, 123.

(14) Ludwig, C.; Gunther, U. L. *BMC Bioinf.* **2011**, *12*, 366.

(15) Hao, J.; Liebeke, M.; Astle, W.; De Iorio, M.; Bundy, J. G.; Ebbels, T. M. *Nat. Protoc.* **2014**, *9*, 1416−1427.

(16) Ravanbakhsh, S.; Liu, P.; Bjorndahl, T. C.; Mandal, R.; Grant, J. R.; Wilson, M.; Eisner, R.; Sinelnikov, I.; Hu, X.; Luchinat, C.; Greiner, R.; Wishart, D. S. *PLoS One* **2015**, *10*, No. e0124219.

(17) Rohnisch, H. E.; Eriksson, J.; Mullner, E.; Agback, P.; Sandstrom, C.; Moazzami, A. A. *Anal. Chem.* **2018**, *90*, 2095−2102.

(18) Lefort, G.; Liaubet, L.; Canlet, C.; Tardivel, P.; Pere, M. C.; Quesnel, H.; Paris, A.; Iannuccelli, N.; Vialaneix, N.; Servien, R. *Bioinformatics* **2019**, *35*, 4356−4363.

(19) Canueto, D.; Gomez, J.; Salek, R. M.; Correig, X.; Canellas, N. *Metabolomics* **2018**, *14*, 24.

(20) Hu, K.; Westler, W. M.; Markley, J. L. *J. Am. Chem. Soc.* **2011**, *133*, 1662−1665.

(21) Bingol, K.; Li, D. W.; Zhang, B.; Brüschweiler, R. *Anal. Chem.* **2016**, *88*, 12411−12418.

(22) Gomez, J.; Brezmes, J.; Mallol, R.; Rodriguez, M. A.; Vinaixa, M.; Salek, R. M.; Correig, X.; Canellas, N. *Anal. Bioanal. Chem.* **2014**, *406*, 7967−7976.

(23) Pereira, B.; Maraschin, M.; Computational Tools for the Analysis of 2D-Nuclear Magnetic Resonance Data. In *Computational Tools for the Analysis of 2D-Nuclear Magnetic Resonance Data*, Springer International Publishing: Cham, 2022; pp. 52−61.

(24) Wilson, S.; Hamilton, M. A.; Hamilton, G. C.; Schumann, M. R.; Stoodley, P. *Appl. Environ. Microbiol.* **2004**, *70*, 5847−5852.

(25) Fuchs, A.; Tripet, B. P.; Ammons, M. C. B.; Copie, V. *Curr. Metabolomics* **2016**, *4*, 141−147.

(26) Leggett, A.; Wang, C.; Li, D. W.; Somogyi, A.; Bruschweiler-Li, L.; Brüschweiler, R. *Methods Enzymol.* **2019**, *615*, 407−422.

(27) Bligh, E. G.; Dyer, W. J. *Can. J. Biochem. Physiol.* **1959**, *37*, 911−917.

(28) Delaglio, F.; Grzesiek, S.; Vuister, G. W.; Zhu, G.; Pfeifer, J.; Bax, A. *J. Biomol. NMR* **1995**, *6*, 277−293.

(29) Li, D. W.; Hansen, A. L.; Yuan, C.; Bruschweiler-Li, L.; Brüschweiler, R. *Nat. Commun.* **2021**, *12*, 5229.

(30) *MestreNova*, 14.0; 2020.

(31) Cobas, C.; Aboutanios, E.; Sykora, S. *Spectrosc. Lett.* **2020**, *53*, 529−535.

(32) Press, W. H. *Numerical recipes in C : the art of scientific computing*. 2nd ed.; Cambridge University Press: Cambridge, New York, 1992; p xxvi, 994 p.

(33) Reynolds, D., Gaussian Mixture Models. In *Encyclopedia of Biometrics*, Li, S. Z.; Jain, A., Eds. Springer: US, Boston, MA, 2009; pp. 659−663.

(34) Dieterle, F.; Ross, A.; Schlotterbeck, G.; Senn, H. *Anal. Chem.* **2006**, *78*, 4281−4290.

(35) Kohl, S. M.; Klein, M. S.; Hochrein, J.; Oefner, P. J.; Spang, R.; Gronwald, W. *Metabolomics* **2012**, *8*, 146−160.

(36) Giraudeau, P.; Tea, I.; Remaud, G. S.; Akoka, S. *J. Pharm. Biomed. Anal.* **2014**, *93*, 3−16.

(37) Van, Q. N.; Issaq, H. J.; Jiang, Q.; Li, Q.; Muschik, G. M.; Waybright, T. J.; Lou, H.; Dean, M.; Uitto, J.; Veenstra, T. D. *J. Proteome Res.* **2008**, *7*, 630−639.

(38) Giraudeau, P. *Magn. Reson. Chem.* **2014**, *52*, 259−272.

(39) Chylla, R. A.; Hu, K.; Ellinger, J. J.; Markley, J. L. *Anal. Chem.* **2011**, *83*, 4871−4880.

(40) Niklasson, M.; Otten, R.; Ahlner, A.; Andresen, C.; Schlagnitweit, J.; Petzold, K.; Lundstrom, P. *J. Biomol. NMR* **2017**, *69*, 93−99.

(41) Smith, A. A. *J. Biomol. NMR* **2017**, *67*, 77−94.

(42) Dudley, J. A.; Park, S.; MacDonald, M. E.; Fetene, E.; Smith, C. A. *J. Magn. Reson.* **2020**, *318*, 106773.

(43) Storey, J. D.; Tibshirani, R. *Proc. Natl. Acad. Sci. U. S. A.* **2003**, *100*, 9440−9445.

(44) Leggett, A.; Li, D.-W.; Sindeldecker, D.; Staats, A.; Rigel, N.; Bruschweiler-Li, L.; Brüschweiler, R.; Stoodley, P. Cadaverine Is a Switch in the Lysine Degradation Pathway in Pseudomonas aeruginosa Biofilm Identified by Untargeted Metabolomics. *Front. Cell. Infect. Microbiol.* **2022**, *12*, DOI: 10.3389/fcimb.2022.833269.